Fusion gene detection using RNAseq

B. Tops, PhD ECP2019
Pediatric oncology in the Netherlands

- ~600 new pts / yr (0 - 18 jaar)
- Many different tumor types: most <25 cases / yr
- 5-yr survival ~75%
- Most frequent cause of death (diseases) among children
- Many late effects (treatment-related)
Why RNAseq (1)

Quick inventory:
54 different 5’ gene fusion partners
59 different 3’ gene fusion partners
Why RNAseq (2)

Care (diagnostics)  

Synergy

Research
Why RNAseq (3)

- Many gene fusions in pediatric cancer
- Fast moving field
- Research hospital (data is shared with pre-clinical research if IC is signed)
RNAseq

- Targeted
- mRNA (polyA enrichment)
- Total RNA (with ribo-depletion)
RNAseq

- Targeted
- mRNA (polyA enrichment)
- Total RNA (with ribo-depletion)

- Input: 50-300 ng total RNA from fresh (frozen) material
- 1 day library prep
- Sequencing on a NovaSeq 6000:
  - Insert size: ~300 bp
  - 2x 150 nt sequencing
  - ~70 mln uniquely aligned reads
Bio-informatics

• In general we follow the best practices from the Broad Institute
• Currently no best practices for fusion transcript detection

• STAR (aligner + fusion):
  • Experience with STAR within institute
  • Easy to implement in workflow
Bio-informatics

Finding seeds

Exact match? yes

Clustering seeds

Stitch seeds

Bio-informatics

Finding seeds

Exact match?

Good extension?

Clustering seeds

Stitch seeds

Finding seeds

Exact match?

yes

no

Good extension?

yes

no

Soft clip seeds

Clustering seeds

Stitch seeds

adapter (or poor quality)
Finding seeds

Exact match?

Good extension?

Soft clip seeds

Clustering seeds

Stitch seeds

Stitched read

donor site
acceptor site

seed1 seed2

read

• Map to gene annotations
• Extra filtering
  • Remove duplicate paired-end alignments
  • Remove weakest supported candidates
• Candidate gene pairs
  • Sense-sense orientation
  • Scoring based on split and spanning reads
• Filter out unlikely candidates
  • Grouping by breakpoint proximity
  • Assessing strength of alignment evidence
  • Filtering lowly supported isoforms
  • Filtering sequence-similar fusion pairs
  • Filtering promiscuous fusion partners
Results

- Test set of 25 samples, containing:

<table>
<thead>
<tr>
<th>Fusion</th>
<th>22/25 fusions detected</th>
<th>1x fusion found with different threshold</th>
<th>1x missed (sub-clonal event)</th>
<th>1x missed, complex rearrangement</th>
</tr>
</thead>
<tbody>
<tr>
<td>NAB2-STAT6</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RBM15-MKL1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>KMT2A-MLLT3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EWSR1-FLI1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SS18-SSX2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FUS-CREB3L2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RUNX1-RUNX1T1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NUP98-KDM5A</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>KMT2A-MLLT1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>KMT2A-MLLT10</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>KMT2A-AFDN</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>KM2TA-AFF1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CBFA2T3-GLIS2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RUNX1-RUNX1T1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BCR-ABL1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RUNX1-ETV6</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CBFB-MYH11</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### Results

- Test set of 25 samples, containing:

<table>
<thead>
<tr>
<th>Fusion</th>
<th>Fusion</th>
<th>Fusion</th>
<th>Fusion</th>
<th>Fusion</th>
</tr>
</thead>
<tbody>
<tr>
<td>NAB2-STAT6</td>
<td>RBM15-MKL1</td>
<td>KMT2A-MLLT3</td>
<td>EWSR1-FLI1</td>
<td>SS18-SSX2</td>
</tr>
<tr>
<td>FUS-CREB3L2</td>
<td>RUNX1-RUNX1T1</td>
<td>NUP98-KDM5A</td>
<td>KMT2A-MLLT1</td>
<td>KMT2A-MLLT10</td>
</tr>
<tr>
<td>KMT2A-AFDN</td>
<td>KM2TA-AFF1</td>
<td>CBFA2T3-GLIS2</td>
<td>RUNX1-RUNX1T1</td>
<td>BCR-ABL1</td>
</tr>
<tr>
<td>RUNX1-ETV6</td>
<td>CBFB-MYH11</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- 22/25 fusions readily detected
  - 1x fusion found with different treshold (FFPM)
  - 1x missed (sub-clonal event)
  - 1x missed, complex rearrangement? **POLD3-KMT2A** and **MLLT10-RPS3**
Results

• After 1\textsuperscript{st} test, filtering adapted:
  • No automatic filtering based on ‘red herrings’ gene fusion list
  • Add white list so genes are not filtered out based on low FFPM values (fusion fragments per million total reads)

• Consequence is quite a lot of manual curation
Results

• Since December 297 samples prospectively sequenced:

  - 60 clinically relevant fusion events detected by RNAseq and ‘classical’ techniques
  - 28 clinically relevant fusion events detected only by RNAseq
  - 1 clinically relevant fusion event only detected by SNP array

50% more relevant fusion transcripts detected using RNAseq
Results

Reasons for additional fusions detected by RNAseq:

1. Not tested for by ‘classical techniques’ (n=18):
   - DNAJB1--PRKACA (hepatocellular carcinoma)
   - ZCCHC8--ROS1 (high grade glioma)
   - EML4--NTRK3 (high grade glioma)
Sensitivity

When is a negative result (un)reliable?
Multiple variables:

- RNA input
- RNA integrity
- Tumor cell content
- Expression levels of fusion genes
RNA concentration (Allprep kit Qiagen)

<5% of the samples not suitable for RNAseq
RNA integrity
RIN vs. reads

No strong correlation between RIN value and uniquely aligned reads
5 FFPE tissues positive for *NTRK* fusions (Heidelberg)

<table>
<thead>
<tr>
<th>ID</th>
<th>Fusion</th>
</tr>
</thead>
<tbody>
<tr>
<td>PMABM000EDD</td>
<td>ETV6—NTRK3</td>
</tr>
<tr>
<td>PMABM000EDE</td>
<td>AGAP1—NTRK2</td>
</tr>
<tr>
<td>PMABM000EDF</td>
<td>KCTD16—NTRK2</td>
</tr>
<tr>
<td>PMABM000EDG</td>
<td>KIF5A—NTRK2</td>
</tr>
<tr>
<td>PMABM000EDH</td>
<td>KCTD8—NTRK2</td>
</tr>
</tbody>
</table>
Possible quality parameters

- $DV_{200}$: the percentage of RNA fragments that are $>200$ nucleotides in size
Conclusions

- RNAseq is robust
Possible future applications

- Identifying druggable targets based on (over)expression
- Tumor classification
